University of Southern Denmark
IMADA
Arthur Zimek
Jonatan Møller Gøttcke, Jonas Herskind Sejr

DM566/DM868/DM870: Data Mining and Machine Learning

Spring term 2019

Exercise 11: Clustering Lab

Exercise 11-1 Clustering Lab-Session

For each of the tasks you can use either ELKI, R, scikit-learn (Python), or another tool of your choice.

(a) Data Preprocessing:

- Go to `https://archive.ics.uci.edu/ml/datasets/seeds` and read about the seeds dataset. Download the dataset.
- The dataset is presented in a tab separated format. The format might not be immediately suitable to analyze the dataset with your preferred analysis tool.

  Try and reformat the dataset for use with the tool of your choice. The ELKI formatted version can be found on e-learn.

(b) k-means:

- Use a k-means implementation in your preferred tool to analyze the seeds dataset.
- What is a suitable choice for k?
- Try different k-means variants like MacQueen, Lloyd, Elkan, k-means++. Do you observe any differences or tendencies in the results?

(c) EM-clustering:

- Run EM clustering on the seeds dataset.
- What is a suitable choice for k?
- For different choices of k compare the result to a similar choice of k in the k-means algorithm.

(d) DBSCAN:

- Run DBSCAN on the seeds dataset.
- Find suitable parameter values for epsilon and minpts.

(e) You might also want to try SNN clustering or hierarchical clustering. Given your experience with this dataset by now – does using these algorithms make sense on this dataset?

(f) Comparison:

Which type of clustering algorithm do you consider the most suitable for this dataset?

(g) If you finish early, try a different tool, and start from the preprocessing step again to get familiarized with that tool.